Assessing the Accuracy of Geocoding Using Address Data from Birth Certificates: New Jersey, 1989 to 1996

Mark C Fulcomer (1),* Matthew M Bastardi (2), Haniya Raza (1), Michael Duffy (1), Ellen Dufficy (1), Marcia M Sass (1)

(1) New Jersey Dept. of Health and Senior Services, Center for Health Statistics, Trenton, NJ; (2) New Jersey Dept. of Treasury, Office of Telecommunications and Information Systems, Trenton, NJ

Abstract

With the widespread availability of low-cost geographic information systems (GIS) on microcomputers, there has been growing interest in linking sociodemographic variables from census and other sources to vital records for individuals (e.g., from birth and death certificates). Such linked data sets would especially assist investigations into factors contributing to adverse reproductive outcomes (e.g., very low birth weight, infant mortality) and other health events in small areas. Address standardization software with built-in geocoding features offers particular promise in appending data from locations such as census tracts. Because successful linkages of social area and individual levels of data rely on accurate geocoding information, this presentation examines the quality of address data from a large, population-based vital records system. Expanding on an earlier report that studied adverse reproductive outcomes for 1985 to 1988, this paper describes New Jersey's efforts to assess the accuracy of locational data reported on its 1989 to 1996 birth certificates (N=971,592) with that resulting from the application of an address standardization procedure (N=951,895). At the municipality level the agreement between geocoding from address standardization and the certificates was 91.68%, while for 870,149 (91.41%) of the records the census tract or block group could be identified. Before the results could be compared, preliminary work of reviewing and correcting some records was required, especially for post office boxes and rural delivery addresses. Because many records fall into areas spanning multiple municipalities, the results will affect linkages of zip code information for municipalities. Specifically, with considerable confusion between zip code and municipality boundaries, methods to minimize misclassification errors will have major implications for projecting school enrollments and estimating health outcomes.

Keywords: address, geocoding, census

Introduction

The major purpose of this paper is to describe the accuracy and utility of birth certificate mailing address data in geocoding municipalities (also known as minor civil divisions, or MCDs) compared with traditional coding of MCDs using mother's residence also listed on the same vital record. Analyses were conducted for the birth years 1989 to 1996, following the implementation of New Jersey's variant of the national standard certificate. This paper introduces a problem that came to the attention of the New

* Mark C Fulcomer, New Jersey Dept. of Health and Senior Services, Center for Health Statistics, Health and Agriculture Bldg, Room 405, Trenton, NJ 08625-0360 USA; (p) 609-984-6702; (f) 609-984-7633; E-mail: mcf@doh.state.nj.us

Jersey Department of Health and Senior Services (NJDHSS) Center for Health Statistics (CHS) early in 1991—that of confusing results arising from different methods used for assigning geocodes. The paper then describes a multi-step process followed to improve the quality of addresses and other information on its birth certificates, including the introduction in 1995 of the most ambitious electronic birth certificate (EBC) system in the entire country (1).

New Jersey, along with several other states, has a long history of reporting births and other health outcomes at the municipality level (2). Typically, these reports are based on statistical analysis of the numeric values (often referred to as geocodes) of such areas as recorded on vital records, with little attention to how accurately the coding process reflects the actual MCDs. Although there has been a growing interest in using sophisticated geographic information systems (GIS) to carefully pinpoint the residences of cases for cluster investigations and other epidemiological studies of possible environmental exposures (3), these efforts have usually focused on relatively small geographic areas such as a few zip codes or census tracts. The emerging capabilities, however, of address standardization software procedures with GIS features has made it feasible and inexpensive to expand the computerized geocoding of events to much larger areas. This report presents some results from what appears to be among the first large-scale, population-based studies (i.e., data from several years for an entire state) comparing the underlying accuracy of traditional manual geocoding of MCDs on vital records with that found automatically through address standardization. While perhaps giving initial impressions of being specific to New Jersey, some of the confusions between address data and traditional geocodes are likely to be encountered in other areas with large populations, especially as software packages for address standardization and GIS are more widely applied.

Background

New Jersey's efforts to improve the quality of its geocoded information on the state's municipalities began early in 1991, shortly after CHS was reorganized under its current director. Because the funding of New Jersey's schools relies heavily on taxes collected and administered at the local level (all 566 of its current MCDs are incorporated and there are 611 distinct school districts), there is considerable interest in projecting school enrollments to estimate future classroom construction needs using cohort survival methods. As a result, representatives of many of the districts call CHS for the latest official birth statistics for selected municipalities and counties in order to develop their projections.

In the process of responding to requests from school planners for the then newly available data from 1989 birth certificates, CHS staff encountered several instances of enormous changes between birth figures for 1988 (and earlier years) and 1989. These shifts appeared to have been due to the introduction of a new birth certificate form for 1989 births. Table 1 shows shifts in birth figures that involved an apparent 241% increase in births between 1988 and 1989 in a small town with a relatively high median age, while there was a corresponding drop of births (-86%) for the same two years in a surrounding municipality that shared the first town's zip code. Although it involved smaller percentage changes in the birth attribution between 1988 and 1989 for two adjacent municipalities (28% and -34%, respectively), the second example shown in

Table 1 Examples of Shifts in Birth Figures Associated with the Introduction of New Certificates in 1989

EXAMPLE #1

		Number of Births Per Year											
MCD	**Code Type**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
1A	Old	72	57	89	99	338	334	90	75	82	89	80	77
	New	—	—	—	—	91	99	63	65	66	72	71	68
1B	Old	130	161	218	203	28	61	274	295	273	254	258	229
	New	—	—	—	—	245	265	295	288	274	251	249	225

EXAMPLE #2

		Number of Births Per Year											
MCD	**Code Type**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
2A	Old	1,888	1,951	2,061	2,136	2,728	2,614	2,624	2,072	1,805	1,663	1,569	1,549
	New	—	—	—	—	2,112	2,015	2,056	1,926	1,777	1,633	1,518	1,521
2B	Old	1,003	1,034	1,081	1,122	740	794	737	1,010	1,103	993	968	1,004
	New	—	—	—	—	1,155	1,190	1,134	1,129	1,121	1,032	1,026	1,025

Old=Traditional vital records geocoding of the mother's residence municipality as reported on the birth certificate
New=Coding of residence municipality based on mother's mailing address as reported on the birth certificate

Table 1 represents an even more dramatic situation, because the first municipality had to account to state government for nearly $9,000,000 in excess payments it had received based on faulty enrollment projections (4).

The fundamental problem that needed to be resolved was the confusion between representing a mother's mailing address as a postal address versus representing the same address relative to the boundaries of the municipality/MCD. This confusion can be visualized by overlaying MCDs within a county on top of zip code boundaries for the postal deliveries to those same areas. Especially vivid are overlays that display zip codes overlapping both municipality and county boundaries, a situation that is not unique to New Jersey.

While participating in the 1989 nationwide implementation of a new standard birth certificate, New Jersey had contributed to the blurring of MCD and zip code boundaries by inadvertently placing the birth certificate query for the mother's mailing address before that for her municipality of residence. Although this seemingly minor reversal was soon corrected in the printed birth certificates (in mid-1991), considerable confusion persisted, in part because the collection of many additional items on a multi-part form (instead of the more compact, 5- by 8-inch version used prior to 1989) had altered the methods used by hospitals to prepare typed versions of the certificates. In particular, the new certificate forms could not easily be copied for use by parents/informants in completing selected items as had been done with the forms for earlier years. Instead, many hospitals began to substitute the postal city of the mother's mailing address for the municipality of residence, often without questioning the mother at all.

Beyond their role in projecting school enrollments and classroom needs, birth figures have had several important health applications reported on by CHS staff and associates. Especially noteworthy are those instances in which live births provide denominators for the calculation of rates for analysis of geographic variations, including low birth weight and inadequate prenatal care, as well as other birth-related characteristics such as infant and fetal mortality (2); case-control studies of environmental exposures and adverse reproductive outcomes (3); and cluster investigations (5). Of course, an even more important consequence of any misclassification of events in adjacent areas is the possibility that the "numerator" characteristics involved in the calculation of such rates may also be affected when the records of some individuals are assigned to different municipalities. For example, those infants who were determined to be inappropriately geocoded to municipality "2A" in Table 1 had a mean birth weight that was approximately 150 grams higher than those births for whom the original geocode was unaltered (6); that is, the misallocation of cases would paint a much more optimistic picture of the health status of infants in "2A" than would be warranted.

In retrospect, given that it is the nation's most densely populated state and has a tradition of local home rule dating back to the American Revolution, New Jersey's geocoding difficulties are hardly surprising. In contrast to the generally rectangular partitioning of other parts of the country that later became part of the emerging nation (e.g., the Northwest Territory), the boundaries of New Jersey's counties and municipalities are irregular and often lack readily identifiable physical demarcations such as rivers or roads. Other confusions stem from instances of duplicate municipality names in different counties, in recognition of important Revolutionary War figures (e.g., 6 instances of Washington Townships, 4 Franklin Townships, etc.). There are also confusions that occur in about 25 pairs of adjacent municipalities (e.g., Princeton Boro and Princeton Township) in which the post office serving a central area also delivers mail to a surrounding MCD with a nearly identical name. Furthermore, with a long and colorful history, New Jersey has about 3,600 small areas that are known by local names, most of which are not municipalities and, therefore, lack any official governmental status or well-defined boundaries, despite their sometimes distinctive-sounding names. Toms River, a part of Dover Township in Ocean County and site of an ongoing childhood cancer cluster investigation, is perhaps the most well-known recent example of these local name areas.

When one considers New Jersey's municipalities and local name areas in conjunction with the postal zip codes serving them, the basis of geocoding confusion becomes even more apparent. Based on the "good" addresses employed in this report—those that met post office certification standards and could be geocoded unambiguously by census TIGER boundaries with no alterations, the state's MCDs are overlapped by 624 zip codes representing 659 different postal city names. Because postal delivery routes are not required to correspond to other geopolitical entities and are sometimes changed to improve service, a large number (392) of the state's zip codes cross municipality boundaries, sometimes even crossing counties in the process. As part of early work on this study, 360 of New Jersey's municipalities were identified as being affected by a substantial sharing of mail delivery routes with neighboring communities, while 42 of its zip codes cross into a second county and 5 of those cross into yet a third county. Although the sharing of zip codes most often affects pairs of municipalities (152 affected), there are instances of zip codes that serve as many as twelve or thirteen different MCDs (2 and 1 affected, respectively). Finally, there were 1,932 combinations of postal cities, zip codes, and municipalities that accounted for the locations of the good records in the present study. (Note that ongoing population growth/migration and the closing of smaller post offices are among factors likely to lead to the future recognition of similar confusions between zip code and MCD boundaries in other states.)

Steps to Improve Quality

Once the scope and nature of the geocoding confusion on New Jersey birth certificates became apparent in early 1991, CHS and NJDHSS's Bureau of Vital Statistics (BVS) began coordinating a series of steps to improve the quality of the state's locational data for these records. Early steps concentrated on the traditional geocoding of municipalities prepared by BVS. Then, after CHS gained access to address data from birth certificates (also based largely on work performed by BVS), efforts shifted toward making such information more useful by geocoding with greater accuracy, not only at the municipality level but also for smaller areas such as census block groups. (Clearly, some of these steps could be replicated elsewhere.)

Hospital Visits

The hospital accounting for the vast majority of the confusion in the first example cited above was visited on three separate occasions. Working with the local registrar in that municipality, hospital records were inspected and birth records were amended to reflect the actual municipalities of residence cited (versus the apparent municipalities reflected in the postal cities listed in mailing addresses). While such a process would be labor-intensive if done on a statewide basis, these initial visits gave valuable insights into subsequent efforts to improve geocoding accuracy.

Change Order of Birth Certificate Items

Simultaneous with the hospital visits, the order of the mailing address and residence municipality items on the birth certificates was reversed. Unfortunately, this change was probably only minimally effective because it was made long after the new data collection methods had been introduced by hospitals to account for the vast increase in information collected on the new 1989 certificates. However, as a new cycle of national standard certificates are introduced in the near future, it is hoped that more attention will be paid to improving and standardizing the acquisition of some key pieces of information such as residential locations and race/ethnicity. A particularly critical change would seem to be the inclusion of direct, personalized probing of sensitive information (versus relying on mailing addresses and visual attributions of race and ethnicity).

Comparison of Statistical Results to Street Maps

By the summer of 1991, CHS had completed the process of inspecting street maps for the state's 21 counties to better understand the changes in geocoding results between 1988 and 1989 births for adjacent municipalities. In the absence of a computerized mapping capability, this step was more difficult than initially envisioned, especially when using out-of-date or incompatible maps to inspect zip codes that crossed county boundaries. As a result of the statistical/map comparison work, however, 360 municipalities were identified as having substantial overlapping of zip codes from nearby communities.

Design and Implementation of Birth Certificate Worksheets

By the fall of 1991, CHS had designed, pilot-tested, and revised four-part worksheets to improve the quality of birth certificate data by standardizing its collection. Central to this effort was a parents' information sheet that concentrated on carefully ascertaining the mother's residence and mailing addresses as well as other items relating to race and ethnicity. The worksheets were implemented statewide through regional training sessions in the spring of 1992 and provided the basis for the eventual introduction of New Jersey's ambitious EBC system in 1995. The second example in Table 1 indicates how the worksheets had an apparent impact on improving the traditional coding of municipalities as early as 1992 and 1993, long before any work on address standardization had commenced.

Interactions with Local Registrars, Hospital Personnel, and School Officials

CHS initiated discussions with and sought feedback from local registrars and hospital personnel to improve the design of the birth certificate worksheets. Being responsible for critical components of the birth certificate process, these groups were seen as key players in a data quality improvement effort. Later, as part of attempts to explain oftentimes large variations in birth figures over time, there were also hundreds of interactions with school superintendents and planners, clearly underscoring the complexities of this project, especially with respect to predicting school enrollments.

Initial Inspections of Addresses for In-State Records

By the winter of 1992, CHS had gained access to computerized files of mothers' mailing addresses on birth certificates. Although it was quickly realized that considerable effort would be required to correct keypunching errors and parse the information into separate fields (e.g., street numbers and names) in order to perform a meaningful analysis, the ability to eventually access improved versions of these data reinforced the efforts to implement the worksheets. Later in 1992, the design of the EBC system began and included an early commitment to standardize the collection of addresses and other important information.

Discovery of Address Standardization with Census Geocoding

The major breakthrough in this project came in the winter of 1994 when, as part of a general interdepartmental discussion with the New Jersey Department of Environmental Protection on how sophisticated GIS techniques might be applied to the birth geocoding problem, it was discovered that a sister New Jersey agency (OTIS, the Office of Telecommunications and Information Systems) could support access to a state-of-the-art address standardization software package, Finalist/FinalFocus. As a tool for achieving valuable postal discounts through the assignment of zip+4 codes, the software would provide important data-cleaning and parsing features. Even more important, a little investigation soon revealed that census tract and block group identifiers were not only employed internally as part of the standardization procedure (along with latitudes and longitudes), but they would also be returned as part of New Jersey's acquisition of the software package. This breakthrough meant that, for the first time, there was the prospect of a computerized procedure that could standardize "messy" address data and automatically assign geocodes to records.

Acquisition of Addresses for 11,509 Out-of-State Births from Pennsylvania for 1989 to 1993

Once the address standardization procedure became available, CHS's attention turned to the acquisition of mailing address information that had previously been missing, beginning with that for out-of-state events. A special arrangement made available an electronic file of addresses for New Jersey resident events occurring in Pennsylvania for the years from 1989 to 1993, thereby eliminating the need to re-key this information.

Data Entry of Birth Addresses for Additional Births

Beginning in the summer of 1994, CHS staff began the process of keying previously missing address information, eventually accounting for the entry of such data for an additional 10,242 in-state births. Note that, prior to this step, mothers' mailing addresses were entered when a social security card was requested for a child, covering 96.28% of the births in New Jersey. For 1991 and 1992, this data entry work was concentrated on records from the 360 municipalities with identifiable geocoding confusion. Unfortunately, because interstate agreements limit how long exchanged vital events information may be retained, out-of-state records for the 1989 and 1990 birth-years were no longer available (except those already supplied by Pennsylvania, as described above), so that no additional address information was initially keyed for those two data years. More recently, all address information for New Jersey births has been computerized; CHS completed this work for the 1993 birth year, while BVS made this part of its routine activities beginning in 1994.

Because of the importance of tracking geographic variations in infant mortality, CHS staff then found and entered previously missing birth certificate address information for 754 infant deaths. Much of these data were missing due to many of these events occurring soon after birth, so that a social security card would never have been requested. In turn, this led to a decision by CHS staff to key address information for the remaining 9,277 in-state events with previously missing computerized data for which a paper certificate could be found, regardless of the geocoded municipality or data year. As a result of keying in previously missing data (much of it from larger cities that were not initially identified as being affected by geocoding confusions), addresses were available for virtually all in-state births for the entire study period from 1989 to 1996.

Initial Results and Corrections/Handling: 1989 to 1994

By the fall of 1996, the initial results for births from the 1989 to 1994 data years became available. Although the standardization process provided helpful guidance on how to inspect and improve addresses (e.g., handling rural delivery routes and post office boxes), these early results were especially encouraging in that they indicated how what had appeared to be very "messy" data could be automatically geocoded in about 90% of the cases. This hands-on experience provided an opportunity to visually inspect 26,459 records in which zip codes were changed by the software to achieve SOUNDEX matching to the street addresses as well as the rendering of what initially appeared to be 9,633 rural delivery routes and 11,638 post office box addresses. Some suggestions on improving the parsing of addresses and the correction of a few census TIGER bugs were passed on to the software vendor and became part of the regular quarterly updates.

Subsequent Software Enhancements

Based on the initial results, major enhancements of the geocoding aspects of the address standardization process were undertaken in the winter of 1997. Instead of leading to an overall latitude/longitude assignment for a census block group to which a good address would have been assigned under the previous version, the enhanced procedure provided interpolated values (over the range in a street segment) for individual records to compare with census TIGER boundaries at the census block level. (The pair of latitude/longitude values for an entire census block group were still inserted when an address could only be matched at the zip+4 level.) To improve address matching through the use of large postal and census databases to augment the Finalist/FinalFocus results, the new procedure also (a) incorporated alternate street names; (b) accounted for street numbers beyond current TIGER file limits (e.g., for newly constructed homes); (c) corrected a "county-road problem" (i.e., by assigning rural deliveries to the county in which the residence was located in those instances where the mail box was on the opposite side of the road and, therefore, in another county); (d) provided more "return codes" to facilitate analysis of the computerized records; and (e) reduced the number of addresses falling into areas spanning multiple municipalities, because census blocks typically do not cross MCD boundaries. (Note that, beyond the initial cost to acquire the Finalist/FinalFocus package, the minimal costs for these enhancements were the only other direct costs for this entire project.) All of the 1989 to 1994 addresses were then re-processed using the new procedure. By the summer of 1998 it was also possible to complete the processing of the 1995 and 1996 certificates.

Acquisition of Addresses for 2,505 Out-of-State Births for 1996

The final quality improvement step involved an acquisition of 1996 addresses for New Jersey residents born in New York City. This step was facilitated by both jurisdictions implementing EBC systems that year using software developed by the same vendor. Much like the earlier acquisition of out-of-state data from Pennsylvania, this step also meant that the addresses for the 2,505 cases could be used directly without re-keying.

Results

This section describes the basic results obtained over the eight years of birth data covered by this entire project, beginning with simple summaries of the acquisition of the address information and efforts to standardize it. After highlighting efforts to attach locational indicators from census data to the standardized address data, the report concludes with comparisons of geocoding accuracy by the two major methods employed.

Acquisition of Address Information

Table 2 shows the results of efforts to acquire (and key) address data from the 971,592 births covered by the eight-year period from 1989 to 1996. Although the project had

Table 2 Births by Data Acquisition Source and Year

	No. of birth records obtained								
Data Source	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**Total**
BVS	115,917	117,228	116,425	115,660	115,146	114,429	117,155	113,866	925,826
PA/NYC	2,377	2,360	2,300	2,214	2,258	0	0	2,505	14,014
CHS/BVS	0	0	971	1,048	2,741	5,482	0	0	10,242
CHS/ID	199	167	202	185	1	0	0	0	754
CHS	2,757	2,733	2,083	1,632	72	0	0	0	9,277
Unavail.	3,401	3,492	2,233	2,321	24	8	0	0	11,479
Total	124,651	125,980	124,214	123,060	120,242	119,919	117,155	116,371	971,592

BVS=Bureau of Vital Statistics (NJ)
PA/NYC=State of Pennsylvania/New York City
CHS=Center for Health Statistics (NJ)
CHS/BVS=Joint effort by CHS and BVS
CHS/ID=CHS-provided statistics on infant deaths; records missing birth certificate address information

started with no address data whatsoever, Table 2 clearly demonstrates that by its conclusion such information had been attached to all except 11,479 (1.18%) of the records. Even more importantly, there is essentially a complete accounting of address data for the last four years.

Of the data sources presented in Table 2, BVS clearly accounted for the vast majority (95.29%) of the addresses. Prior to the pilot testing of the EBC (at four hospitals in 1995) and its statewide implementation beginning in 1996, the data entry of addresses by BVS was done for those infants for whom social security cards had been requested. Note that the increase in 1995 reflected BVS's keying of addresses for all births prior to the full implementation of the EBC, a process completed by April 1997, so that almost all address information is now provided directly (i.e., without any data entry by BVS) by the 71 birthing facilities in the state.

The number of addresses acquired from Pennsylvania and New York is shown in Row 2 of Table 2 and, at first glance, would appear to represent only a small portion (1.44%) of the total. However, beyond saving the effort involved in re-keying records for out-of-state events, such interstate exchanges hold great promise for improving the timeliness with which population-based vital statistics become available in the future, especially if this data sharing can be expanded to include all states and other data (e.g., death certificates).

Joint efforts by CHS and BVS to key some missing records are highlighted in Row 3 of Table 2. This work began with CHS inputting data for 1991 and 1992 certificates from those municipalities with already identified geocoding problems and was carried over to all remaining records in 1993 and beyond (by CHS and BVS, respectively).

Row 4 of Table 2 lists the number of records entered by accounting for the 754 infant deaths with previously missing birth certificate address information (i.e., in addition to those infant deaths with data already summarized in the first three rows). Because virtually all addresses were keyed after 1992, all except one of these cases occurred between 1989 and 1992. Similarly, CHS's efforts to account for the 9,277 remaining in-state records with previously missing addresses are shown in Row 5 of Table 2.

Finally, Row 6 of Table 2 displays the number of records that were no longer available to provide any address information, primarily from those out-of-state events in 1989 to 1993 that did not occur in Pennsylvania (i.e., those records already accounted for in Row 2). Most of the unavailable addresses were from births that occurred in New York City and came from densely populated areas of Bergen and Hudson counties, where geocoding confusion was not as severe as in other areas of the state.

Address Standardization and Matching

Once data entry work was completed for the 960,113 births for which address information was available, the records were assembled into smaller files and transmitted to OTIS for initial processing. The results of the address standardization procedure were then used to separate records into three major groupings: (1) those that could be matched automatically to a known address and would require no further work (i.e., so-called "good" results); (2) those that could be matched automatically, but only after the Finalist/FinalFocus portion of the procedure changed an address using SOUNDEX matching or other five-digit codes within a three-digit zip code area; and (3) those that could not be matched to an address automatically (i.e., so-called "bad" results). CHS inspected all records in the second group and compared the address matching results after the changes were made with the original information. In those cases in which the changes were not considered acceptable, an attempt was made to edit the address on a record and mark it for resubmission, especially because rather obvious errors (e.g., the failure to join together the number and letter of an apartment such as "2C", leaving it instead as "2 C") can lead to some rather surprising matches (e.g., "2 C Street"). Although the records with bad addresses represented only a small portion of the total, they were too numerous and complex to allow editing to be completed effectively. Instead, those bad records for which there was substantial agreement between the postal city, zip code, and geocode for a given MCD (i.e., when compared with good records for the same area that could be matched to known addresses and geocoded automatically with no alterations) were separated from those that needed further inspection and editing. (Note that some records with out-of-range but consistent street numbers now had interpolated latitude/longitude values attached to them by the new procedure, making them equivalent to those with good results.) Any records marked for editing, whether initially identified by the software as changes or as bad results that could not be matched to a zip+4 area, were then prepared for resubmission to OTIS and the process was repeated with much smaller files. Any records that did not produce acceptable matching after a second attempt were then treated as bad (or problematic) results.

Overall, good matches to known addresses with no alterations were returned by the new procedure for 857,261 (88.23%) of the records. For 39,196 (4.03%) records, the software indicated that some changes were needed to match an address (not all of them were accepted by CHS), while 63,656 (6.55%) were initially identified as unmatchable. In large part, the success in achieving good matches was due to the large number of records (934,746, or 96.21%) viewed as having conventional addresses, including many of which that had been treated as rural deliveries in the initial processing done in 1996.

Rural deliveries (10,224, or 1.05%) and post office boxes (14,946, or 1.54%) accounted for all but 197 of the remaining records with address data.

Attaching Locational Indicators from Census Data

Based on the results of its standardization and matching steps, the new OTIS procedure also attached census identifiers to the 951,895 records with New Jersey addresses (or that were originally coded as in-state residents by BVS when addresses were unavailable). These records are summarized later in the first four rows of Table 3. The remaining 19,697 (2.03%) records had addresses outside of New Jersey (or were coded as out-of-state residents by BVS when addresses were unavailable) and are listed later in Rows 5 and 6 of Table 3. The census locational indicators included tracts, block groups, and blocks. Latitudes and longitudes associated with an address were also attached to the records. With the new OTIS procedure, interpolated latitude and longitude values could be found for many of these records using the census TIGER file limits based on block-level matching. The attachment of a single pair of values for an entire census block group (i.e., geocoding based on the earlier Finalist/FinalFocus procedure) now took place when a census block could not be assigned and an address could be matched only at the zip+4 level.

Table 3 Births by Level of Accuracy and Year of Birth

Level of Accuracy	1989	1990	1991	1992	1993	1994	1995	1996	Total
1 Same MCD	109,902 88.17%	111,510 88.51%	110,705 89.12%	111,489 90.60%	108,783 90.47%	108,126 90.17%	105,799 90.31%	106,400 91.43%	872,714 89.82%
2 Same county	10,169 8.16%	10,215 8.11%	9,560 7.70%	7,835 6.37%	7,917 6.58%	8,317 6.94%	7,749 6.61%	6,520 5.60%	68,282 7.03%
3 Diff. county	1,569 1.26%	1,325 1.05%	1,167 0.94%	1,173 0.95%	1,136 0.94%	1,233 1.03%	1,366 1.17%	1,252 1.08%	10,221 1.05%
4 OOS to NJ	63	68	103	56	75	74	139	100	678
5 NJ to OOS	2	3	2	1	4	8	0	0	20
6 Both OOS	2,946 2.36%	2,859 2.27%	2,677 2.16%	2,506 2.04%	2,327 1.94%	2,161 1.80%	2,102 1.79%	2,099 1.80%	19,677 2.03%

OOS=Out-of-state

After the new procedure matched an address, census locational codes at the tract, block group, block, and other levels were also linked to the records, including a county code and one or more MCD codes. The census values for counties and municipalities have a one-to-one correspondence with the BVS geocodes for the same areas and, therefore, the two sets of geocodes can be used interchangeably. In contrast, because they can cross one or more municipality boundaries (but not county boundaries), census tracts and block groups (i.e., subsets of tracts) may sometimes be shared among multiple MCDs. Fortunately, blocks are generally associated with a single municipality.

Given that it had not even been considered possible at the outset of the project, the geocoding of records to census tracts and block groups has been extraordinarily successful. Of the records for New Jersey residents, 848,189 (89.11%) could be coded to the block group level (which includes tracts) and an additional 21,960 (2.31%) to the tract level only. This result is especially important in that it makes possible the attachment of income and other sociodemographic indicators from social areas, otherwise missing from individual-level records such as birth certificates, in "semi-ecologic" studies of adverse reproductive and other health outcomes (2) or other social area analyses (7).

The use of census locational identifiers, done in conjunction with ranges of street addresses, was also very successful in geocoding records at the municipality level, especially because all except nine of New Jersey's 566 MCDs have boundaries in the TIGER files. With respect to single municipalities, 855,286 (89.85%) of the records for New Jersey residents came from census designation areas that did not cross into another MCD and were mostly geocoded at the block level, except for 105,083 records at the block group level only. For those records that could be geocoded by census indicators but fell into block groups spanning two municipalities (10,169) or three (1,334), the original BVS geocode was relied on as much as possible. Thus, when the BVS geocode derived from the official birth certificate matched one of the possible MCD codes based on the census identifiers, regardless of its position as a member of a pair or triplet, that municipality code was used. This was done for 7,511 of the pairs and 455 of the triplets. Of the remaining 3,537 records falling into census block groups spanning multiple MCDs that did not have a matching BVS code, the municipality codes based on the census identifiers were randomly assigned (using the codes' sequential position in pairs or triplets of possible MCDs for series of similarly sorted records). Again, when contrasted with the early stages of the project, the ability of the address standardization procedure to automatically assign a municipality code to 91.06% of the records in a rational fashion is noteworthy. This result also supports the use of GIS software to display maps of birth figures at the census tract and block group levels, provided that careful attention is paid to protecting the confidentiality of individuals when the number of events in a submunicipality area is small.

Partial matches to the BVS municipal code were established for 67,822 (82.97%) of the 81,746 records for New Jersey residents that could not be automatically matched to census tracts or block groups. This was done using postal cities and zip codes for the same municipalities found in the good results coded to single municipalities. (In the future, these partial matches will be handled through AUTOMATCH [a procedure described in Jaro 1989 (8)], which should improve the geocoding process even more.) As a consequence, only a small number of records (13,924, including those for which vital events information was no longer available, as mentioned earlier) lacked similar concordance between the geocodes from traditional methods and those based on matching address information to huge postal and census databases. For this set of records, only the original BVS geocodes could be used.

Comparing the Agreement between Geocoding Methods

This section describes the agreement between the traditional coding of MCDs of mothers' residences with that based on standardization and matching from address information. Table 3 shows six levels of accuracy used to assess the agreement across the eight-year period from 1989 to 1996. Row 1 highlights the high overall accuracy (89.82% of all records; 91.68% of New Jersey residents) between the two geocoding methods in assigning records to the same municipality. In general, same-municipality agreement has improved over time, especially in 1996 as the EBC was being implemented.

Discrepancies that resulted from the two methods assigning geocodes to different municipalities within the same county are listed in Row 2 of Table 3. The percentage of within-county discrepancies has also diminished over time, perhaps reflecting both the early introduction of the worksheets as well as special features of the EBC software (e.g., pull-down lists of municipalities within counties) because the improvements were most pronounced in 1992 and 1996. An improvement in 1995-1996 data with respect to matching addresses at the census tract or block group levels (95.97% of New Jersey residents versus the overall eight-year average of 91.41%) also likely traces its origins to the better acquisition and keying methods introduced at the birthing facilities as part of the EBC. Taken together, Rows 1 and 2 of Table 3 indicate a relatively high level of accuracy of geocodes at the county level (96.85% of all records; 98.86% of New Jersey residents).

Row 3 of Table 3 shows the number of discrepancies between the two geocoding methods in assigning records to different municipalities in different counties. Unfortunately, the overall percentage of discrepancies between counties is substantial (1.05% of all records; 1.07% of New Jersey residents) and has remained essentially unchanged over the entire eight-year period. Further work, including continued emphasis on increasing the interstate exchange of data, will certainly be needed to understand how and where such errors occur and how they might be ameliorated in the future.

Row 4 of Table 3 shows the number of discrepancies that were originally given out-of-state codes by BVS but which were geocoded as in-state residents using the address data. This relatively minor level of disagreement seems to occur most frequently with births in military families, where permanent homes may be in another state while the use of schools and other resources happens in New Jersey. Row 5 shows the extremely small number of discrepancies that were geocoded as being residents of other states based on the address data but had originally been treated as New Jersey residents by BVS. Finally, Row 6 lists the agreement (2.03% of all records) between the two methods in geocoding records to other states.

Summary

Because it involved nearly one million births over the eight-year period from 1989 to 1996, the entire address standardization/analysis effort was a large and complex undertaking. The ability, however, to successfully resolve what was a great deal of initial confusion has been very gratifying, especially given that the methods can be used elsewhere and that the results have some important applications that were not envisioned at the outset. In particular, the efforts to improve data quality, link records to other data sources (e.g., income from the census), and achieve more timely and automatic geocoding hold great promise for the future.

Nonetheless, despite the clear-cut benefits of automatic geocoding based on the application of standardization and matching techniques to address data, the relatively high disagreements with traditional manual methods at the municipality level are disturbing. The fact that the discrepancies are so large (i.e., regardless of whether they occur in the same or different counties)—nearly 7% even for the most recent year (1996) with the EBC undergoing its full implementation—casts a cloud over the exclusive use of manual methods to geocode residence locations, especially in situations in which address data might be available to facilitate comparisons with results from automatic geocoding. A frightening possibility is how easily "faulty" numerators or denominators from manual geocoding could be employed in the calculation of rates for infant mortality (or other "rare" outcomes) in small areas. Thus, while address data can clearly be helpful in assigning events to small areas, much more work on understanding and improving geocoding methods remains to be done.

References

1. Knoblauch KL, Sherel HP. 1998. *New Jersey electronic birth certificate perinatal database data dictionary*. New Jersey Department of Health and Senior Services, Center for Health Statistics, Trenton, NJ.
2. Fulcomer MC, Bove FJ, Klotz JB, Siegel B, Martin RM. 1992. Developing and utilizing "community health profiles" based on linked information on adverse reproductive outcomes. *Proceedings of the 1991 conference on records and statistics.* US Department of Health and Human Services. DHHS Publication No. (PHS) 92-1214. 168–173.
3. Bove FJ, Fulcomer MC, Klotz JB, Esmart J, Dufficy EM, Savrin JE. 1995. Public drinking water contamination and birth outcomes. *American Journal of Epidemiology* 141:850–62.
4. Bewley J. 1996. City faces $9M loss in school aid. The Trenton *Times.* 5 January. A1, A10.
5. Fulcomer MC, Ziskin LZ, France DM, Bove F. 1988. *Report on the study of Vernon Township, NJ: Study of the occurrence of chromosomal anomalies in Vernon Township between January 1, 1975 and June 30, 1987*. New Jersey Department of Health, Division of Community Health Services, Trenton, NJ.
6. Raza H. 1997. *An analysis of the effects of address standardization of live birth certificates on the health and social characteristics of mothers and babies during 1989-1994 in Trenton, New Jersey.* Unpublished master's thesis. University of Medicine and Dentistry of New Jersey, Piscataway, NJ.
7. Struening EL. 1975. Social area analysis as a method of evaluation. In *The Handbook of evaluation research, Volume I.* Ed. EL Struening and M Guttentag. Beverly Hills: Sage Publications. 519–36.
8. Jaro MA. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* 84:414–20.